Discrete Applied Mathematics 160 (2012) 2339–2348

Contents lists available at SciVerse ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Compact MILP models for optimal and Pareto-optimal LAD patterns☆

Cui Guo [a], Hong Seo Ryoo [b,*]

[a] Graduate School of Information Management & Security, Korea University, 1, 5-Ga, Anam-Dong, Seongbuk-Gu, Seoul, 136-713, Republic of Korea

[b] School of Industrial Management Engineering, Korea University, 1, 5-Ga, Anam-Dong, Seongbuk-Gu, Seoul, 136-713, Republic of Korea

ARTICLE INFO

Article history:
Received 25 April 2011
Received in revised form 16 April 2012
Accepted 17 May 2012
Available online 9 June 2012

Keywords:
LAD
MILP
Strong prime pattern
Strong spanned pattern
Maximum prime pattern
Maximum spanned pattern

ABSTRACT

This paper develops MILP models for various optimal and Pareto-optimal LAD patterns that involve at most $2n$ 0–1 decision variables, where n is the number of support features for the data under analysis, which usually is small. Noting that the previous MILP pattern generation models are defined in $2n + m$ 0–1 variables, where m is the number of observations in the dataset with $m \gg n$ in general, the new models are expected to generate useful LAD patterns more efficiently. With experiments on six well-studied machine learning datasets, we first demonstrate the efficiency of the new MILP models and next use them to show different utilities of strong prime patterns and strong spanned patterns in enhancing the overall classification accuracy of a LAD decision theory.

1. Introduction

Binary classification is a classical problem in data mining and machine learning that deals with the discrimination of two types of data/observations. Supervised learning to binary classification aims to discover a classification/decision theory on past observations to classify new ones in a manner consistent with the past classifications. The Logical Analysis of Data (LAD) is a supervised learning methodology that is based on Boolean logic, combinatorics and optimization [6,7]. A typical implementation of LAD analyzes data on hand via four sequential stages of data binarization, support feature selection, pattern generation and LAD decision theory formation. Here, a LAD decision theory refers to a partially-defined Boolean function built on past observations for classifying new observations.

Let us assume (through the application of the first two stages of LAD, if necessary) that the data under analysis are represented by a small number of 0–1 Boolean variables, called support features. Let us refer to the two types of data as $+$ and $-$ observations and assume that they are contradiction-free such that a LAD decision theory exists for their classification. For $\bullet \in \{+, -\}$, let $\bar{\bullet}$ denote the complementary type of $\bullet$ with respect to the set $\{+, -\}$. A $\bullet$ pattern is a conjunction of one or more literals (where a literal refers to a 0–1 Boolean variable or its negation) that distinguishes at least one $\bullet$ observation from all $\bar{\bullet}$ observations, and the number of literals included in a pattern is called the degree of the pattern. Patterns are the building blocks of a LAD decision theory, and, owing to the significance, the pattern generation has become a central issue in LAD research. However, finding an optimal pattern with respect to a certain pattern preference

☆ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant Number: 2010-0016571).

* Corresponding author. Tel.: +82 2 3290 3394; fax: +82 2 929 5888.
E-mail addresses: guocui@korea.ac.kr (C. Guo), hsryoo@korea.ac.kr (H.S. Ryoo).

0166-218X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2012.05.006

criterion is a difficult combinatorial optimization problem, and most pattern generation methods in the literature are term enumeration-based techniques [2,3,5,7,9].

As seen from the definition of a pattern, the number of patterns of degree d can be as many as $2^d C_d^n$, where n is the number of features. Therefore, the term-enumerative methods can be quite limited in generating 'useful' patterns that are optimal or Pareto-optimal with respect to one or more pattern selection preferences, respectively. To alleviate difficulties associated with generating useful patterns, [12] introduced a Mixed 0–1 Integer and Linear Programming (MILP)-based framework and presented MILP models for generating various optimal and Pareto-optimal patterns for LAD. In short, there are two main advantages of using the MILP approach over the term-enumerative methods. The first is that it generates optimal patterns of different degrees with equal ease (or difficulty) and, in general, without total enumeration. The other is the development of fast algorithms and software for MILPs nowadays. Therefore, the MILP approach can identify useful LAD patterns more efficiently, and [12] demonstrated this point with extensive numerical experiments.

Let us recall basic definitions about LAD patterns. For $\bullet \in \{+, -\}$, a pattern is called a strong $\bullet$ pattern if its coverage is maximum among all $\bullet$ patterns, where the coverage of a pattern refers to the number of observations of a given type the pattern distinguishes from those of the other type. As seen, a strong pattern is optimal with respect to the coverage (or the evidential) preference. For two patterns p_1 and p_2, p_1 is simplicity-wise preferred to p_2 if the literals of p_1 form a subset of the literals of p_2. A pattern is called prime if the removal of any of its literals makes it a non-pattern. A prime pattern is, thus, optimal with respect to the simplicity preference. A pattern can be understood as an interval in $\mathbb{R}^n$, formed by the intersection of the point (attribute) or level variables corresponding to the literals that make up the pattern. A pattern p_1 is selectivity-wise preferred to a pattern p_2 if the interval of p_1 is contained inside the interval of p_2. A selectivity-wise optimal pattern is called a spanned pattern. A strong pattern that is also Pareto-optimal with respect to the simplicity or selectivity criterion is called a strong prime or strong spanned pattern, respectively. Last, a maximum A_i-pattern is a $\bullet$ pattern with the maximum coverage among those patterns that cover a (reference) $\bullet$ observation A_i.

Supervised learning theory basically assumes homogeneity between the past and future observations. This, in turn, suggests that patterns that are optimal or Pareto-optimal with respect to the evidential preference are useful for LAD. In this paper, we consider the evidential preference for LAD patterns and develop an MILP model for generating strong $\bullet$ patterns in Section 2.1. In comparison with the MILP model for strong patterns from [12], the new MILP model involves only $2n$ 0–1 decision variables as opposed to $2n + m^\bullet$, where n is much smaller than $m^\bullet$, the number of $\bullet$ observations. This allows the new MILP model to generate strong patterns more efficiently than its counterpart from [12]. Using this model, we next develop an MILP model for the maximum $C^\bullet(p)$-pattern, which is defined as a $\bullet$ pattern of the maximum coverage among all $\bullet$ patterns that cover a set $C^\bullet(p)$ of $\bullet$ observations. As the reader may see, the maximum $C^\bullet(p)$-pattern of this paper subsumes the maximum A_i-pattern of [5,12] and is the pattern that is called the strong $\bullet$ pattern in the LAD literature [1,2]. After the development of the two MILP models based on the evidential preference, we incorporate the simplicity and selectivity measures and develop four MILP models for strong prime patterns, strong spanned patterns, maximum prime $C^\bullet(p)$-patterns and maximum spanned $C^\bullet(p)$-patterns in Section 2.2. Again, these models are much more compact in terms of hard decision variables than their counterparts from [12]; specifically, they involve $2n$, $2n$, at most n, and at most n 0–1 variables, respectively.

With experiments on six well-studied machine learning datasets, we demonstrate the efficiency of the compact MILP models of this paper over the models we presented in [12] in Section 3.1 and investigate different utilities of strong prime patterns and strong spanned patterns in enhancing the overall classification accuracy of a LAD decision theory in Section 3.2. In supervised learning, a simpler rule is expected to generalize better and classify new observations more accurately (e.g., [4, 10].). As the simplest among all strong patterns, strong prime patterns are thus expected to enhance the classification accuracy of a LAD decision theory. A strong spanned pattern, on the other hand, is the most complex and specific among all strong patterns. Therefore, they are expected to reduce the number of misclassified decisions by a LAD decision theory. In short, numerical results in Section 3.2 strongly support these beliefs; namely that, when filtered through a LAD decision theory, strong prime patterns increase the sensitivity and specificity of decisions and decrease the number of no decisions while strong spanned patterns reduce the number of misclassified decisions.

2. Main results

Let $S^\bullet$ denote the index set of $m^\bullet$ observations of type $\bullet \in \{+, -\}$, and let observation A_i, $i \in S^\bullet$, be described by n 0–1 attributes (support features) $a_j \in \{0, 1\}, j \in N := \{1, \ldots, n\}$. Let a_{ij} denote the binary value the j-th attribute takes on A_i. Let us introduce n additional variables a_{n+j} to negate attributes $a_j, j \in N$; that is, $a_{i,n+j} = \bar{a}_{ij} = 1 - a_{ij}$ for $j \in N$. Let $\mathcal{N} := \{1, \ldots, 2n\}$.

A term t in Boolean logic is a conjunction of literals that can be defined as $t := \bigwedge_{j \in N_t} a_j$ for some $N_t \subset \mathcal{N}$, provided that each $j \in N_t$ corresponds to the index of only one of a_j or a_{n+j}, $j \in N$. In this paper, we say that observation A_i is covered by term t if $t(A_i) := \bigwedge_{j \in N_t} a_{ij} = 1$. Given a term, recall that $C^\bullet(t)$ denotes the index set of those $\bullet$ observations that are covered by term t for $\bullet \in \{+, -\}$. Last, a term is called a $\bullet$ pattern if it distinguishes at least one $\bullet$ observation from all $\bar{\bullet}$ observations. Therefore, a $\bullet$ pattern t satisfies the property that $C^\bullet(t) \neq \emptyset$ and $C^{\bar{\bullet}}(t) = \emptyset$. The coverage of a $\bullet$ pattern is the number of $\bullet$ observations it covers; that is, $|C^\bullet(t)|$. For reasons of space, we refer readers interested in obtaining more background on LAD to [1,7].

2.1. Basic models for strong and maximum patterns

Recall that a strong • pattern is a pattern that distinguishes the largest number of observations in $S^\bullet$ from all those in $S^{\bar\bullet}$. Let $\alpha_{ij} := a_{ij} - 1$ for $i \in S^\bullet$ and $j \in \mathcal{N}$, and let $\beta_{ij} := 1 - a_{ij}$ for $i \in S^{\bar\bullet}$ and $j \in \mathcal{N}$. Consider the following MILP pattern generation model, named $(\mathrm{M}_1^\bullet)$.

$$c_1 = \max_{\mathbf{x},\mathbf{y}} \sum_{i \in S^\bullet} y_i$$

$$\text{s.t. } y_i - \alpha_{ij} x_j \le 1, \quad i \in S^\bullet,\ j \in \mathcal{N} \tag{1}$$

$$\sum_{j \in \mathcal{N}} \beta_{ij} x_j \ge 1, \quad i \in S^{\bar\bullet} \tag{2}$$

$$x_j + x_{n+j} \le 1, \quad j \in N \tag{3}$$

$$\mathbf{x} \in \{0,1\}^{2n} \tag{4}$$

$$\mathbf{0} \le \mathbf{y} \le \mathbf{1} \tag{5}$$

As seen shortly, in a feasible solution to $(\mathrm{M}_1^\bullet)$, $x_j = 1$ indicates that attribute $a_j, j \in \mathcal{N}$, is used in a pattern to be found, and $y_i > 0$ indicates that observation $A_i, i \in S^\bullet$, is covered by the pattern.

Lemma 1. *Let $(\mathbf{x}, \mathbf{y})$ be a feasible solution of $(\mathrm{M}_1^\bullet)$ with $c_1 > 0$. Then, a term defined as*

$$p := \bigwedge_{x_j = 1, j \in \mathcal{N}} a_j \tag{6}$$

forms a • pattern. Furthermore, if $(\mathbf{x}^*, \mathbf{y}^*)$ is an optimal solution of $(\mathrm{M}_1^\bullet)$, then p defined as (6) forms a strong • pattern with coverage c_1^*.

Proof. First, consider $A_l, l \in S^\bullet$, and set $x_j = 1$ if $a_{lj} = 1$ and $x_j = 0$ otherwise; equivalently, set $x_j = 1$ if $\alpha_{lj} = 0$ and $x_j = 0$ if $\alpha_{lj} = -1$. Set $y_l = 1$ and $y_i = 0$ for each $i \in S^\bullet \setminus \{l\}$. The resulting solution $(\mathbf{x}, \mathbf{y})$ now satisfies the constraints in (3)–(5) and has $c_1 = \sum_{i \in S^\bullet} y_i = 1$. Now, note that this solution satisfies $\alpha_{lj} x_j = (a_{lj} - 1) a_{lj} = 0$, hence we have $y_l = \alpha_{lj} x_j + 1$ for $j \in \mathcal{N}$. For $i \in S^\bullet \setminus \{l\}$ with $y_i = 0$, we have $\alpha_{ij} x_j + 1 \ge 0$ (since $\alpha_{ij} \in \{-1, 0\}$), hence $\alpha_{ij} x_j + 1 \ge y_i$ for every $j \in \mathcal{N}$. This shows that $(\mathbf{x}, \mathbf{y})$ satisfies the constraints in (1). Last, for $i \in S^{\bar\bullet}$, there exists $k \in \mathcal{N}$ such that $a_{ik} = 0$ while $a_{lk} = 1$ for $l \in S^\bullet$. So, we have $\beta_{ik} x_k = (1 - a_{ik}) x_k = (1 - a_{ik}) a_{lk} = 1$, and this shows that $(\mathbf{x}, \mathbf{y})$ satisfies the constraints in (2) as well. Therefore, we have shown that $(\mathrm{M}_1^\bullet)$ has at least one feasible solution with $c_1 \ge 1$.

Next, in a feasible solution $(\mathbf{x}, \mathbf{y})$ to $(\mathrm{M}_1^\bullet)$ with $c_1 > 0$, note that there exists at least one $y_i > 0$ for $i \in S^\bullet$. Let $N_p := \{j \in \mathcal{N} : x_j = 1\}$ and form a term by (6) as $p := \bigwedge_{j \in N_p} a_j$. Then, for each $i \in S^\bullet$ with $y_i > 0$, (1) yields $\alpha_{ij} = 0$ for all $j \in N_p$. Therefore, we have

$$C^\bullet(p) = \{i \in S^\bullet : \alpha_{ij} = 0,\ \forall j \in N_p\} \ne \emptyset.$$

Now, let $C^{\bar\bullet}(p) = \{i \in S^{\bar\bullet} : \beta_{ij} = 0, \forall j \in N_p\}$ and note that

$$S^{\bar\bullet} \setminus C^{\bar\bullet}(p) = \{i \in S^{\bar\bullet} : \beta_{ij} = 1 \text{ for some } j \in N_p\} = S^{\bar\bullet}.$$

From these, one can see that the cover inequalities in (2) ensure that every $i \in S^{\bar\bullet}$ belongs to $S^{\bar\bullet} \setminus C^{\bar\bullet}(p)$.

Last, note that the optimization principle will set $y_i = 1$ for each $i \in C^\bullet(p)$ and set $y_i = 0$ for each $i \notin C^\bullet(p)$. The rest of the proof is immediate. □

Let us delete (3) from $(\mathrm{M}_1^\bullet)$ to obtain an MILP model called $(\mathrm{M}_2^\bullet)$ below:

$$c_2 = \max_{\mathbf{x},\mathbf{y}} \sum_{i \in S^\bullet} y_i$$

$$\text{s.t. (1), (2), (4), and (5)}$$

Lemma 2. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be an optimal solution of $(\mathrm{M}_2^\bullet)$. Then, p defined as (6) forms a strong • pattern with coverage c_2^*.*

Proof. Suppose that there exits $k \in N$ such that $x_k^* = x_{n+k}^* = 1$. Then, from (1), we have $y_i^* \le \alpha_{ik} x_k^* + 1$ and $y_i^* \le \alpha_{i,n+k} x_{n+k}^* + 1$ for each $i \in S^\bullet$, where $\alpha_{ik} \in \{0, -1\}$ and $\alpha_{i,n+k}$ takes the complementary value of α_{ik} with respect to the set $\{0, -1\}$. This implies that $y_i^* = 0$ for all $i \in S^\bullet$ and, hence, $c_2^* = 0$. Now, note that $(\mathrm{M}_2^\bullet)$ admits a larger number of feasible solutions than $(\mathrm{M}_1^\bullet)$, whose optimum $c_1^* > 0$. These contradict each other, and this contradicts that $(\mathbf{x}^*, \mathbf{y}^*)$ is an optimal solution of $(\mathrm{M}_2^\bullet)$. □

An interesting property about an optimal solution of $(\mathrm{M}_1^\bullet)$ and $(\mathrm{M}_2^\bullet)$ is summarized in the following proposition.

Proposition 3. *For generating a strong • pattern, (1) in ($M_1^\bullet$) and ($M_2^\bullet$) can be replaced by*

$$y_i - (\alpha_{ij}x_j + \alpha_{i,n+j}x_{n+j}) \le 1, \quad i \in S^\bullet,\ j \in N.$$

Proof. Note that (1) can be written as

$$y_i \le \min\{\alpha_{ij}x_j, \alpha_{i,n+j}x_{n+j}\} + 1 \quad \text{for } i \in S^\bullet,\ j \in N.$$

So, it suffices to show that

$$\alpha_{ij}x_j + \alpha_{i,n+j}x_{n+j} = \min\{\alpha_{ij}x_j, \alpha_{i,n+j}x_{n+j}\} \quad \text{for } i \in S^\bullet,\ j \in N,$$

and this is an immediate consequence of (4) with the property that $x_j + x_{n+j} \le 1$ for $j \in N$ in an optimal solution of ($M_1^\bullet$) and ($M_2^\bullet$). □

The results above yield the following MILP model for strong • patterns. Let us call this model ($M_s^\bullet$):

$$c_s = \max_{\mathbf{x},\mathbf{y}} \sum_{i \in S^\bullet} y_i$$

$$\text{s.t. } y_i - \alpha_{ij}x_j - \alpha_{i,n+j}x_{n+j} \le 1, \quad i \in S^\bullet,\ j \in N \tag{7}$$

$$\sum_{j \in \mathcal{N}} \beta_{ij}x_j \ge 1, \quad i \in S^{\bar\bullet} \tag{8}$$

$$\mathbf{x} \in \{0,1\}^{2n} \tag{9}$$

$$\mathbf{0} \le \mathbf{y} \le \mathbf{1} \tag{10}$$

The following is immediate.

Theorem 4. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be an optimal solution of ($M_s^\bullet$). Then, p defined as (6) forms a strong • pattern with coverage c_s^*.*

Recall that we call a pattern a maximum $C^\bullet(p)$-pattern if it has the maximum coverage among the • patterns that cover all observations A_i, $i \in C^\bullet(p) \subseteq S^\bullet$. To generate $C^\bullet(p)$-maximum patterns, let

$$J_p = \{j \in \mathcal{N} : a_{ij} = 1, \forall i \in C^\bullet(p)\},$$

and consider the following MILP model called ($M_m^\bullet$).

$$c_m = \max_{\mathbf{x},\mathbf{y}} \sum_{i \in S^\bullet \setminus C^\bullet(p)} y_i$$

$$\text{s.t. } y_i \le \alpha_{ij}x_j + 1, \quad i \in S^\bullet \setminus C^\bullet(p),\ j \in J_p \tag{11}$$

$$\sum_{j \in J_p} \beta_{ij}x_j \ge 1, \quad i \in S^{\bar\bullet} \tag{12}$$

$$\mathbf{x} \in \{0,1\}^{|J_p|} \tag{13}$$

$$\mathbf{0} \le \mathbf{y} \le \mathbf{1} \tag{14}$$

The following is immediate.

Theorem 5. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be an optimal solution of ($M_m^\bullet$). Then, p defined as*

$$p^\bullet = \bigwedge_{j \in \{J_p : x_j = 1\}} a_j$$

forms a maximum $C^\bullet(p)$-pattern with coverage $c_m^* + |C^\bullet(p)|$.

We give three remarks here.

Remark 6. Despite the fact that they are developed independently, there is a similarity in looks between the MILP model for strong patterns from [12], called (MILP-1•), and ($M_2^\bullet$) above. To help in understanding, let us recall the formulation of (MILP-1•) below:

$$z = \min_{\mathbf{x},\mathbf{w},d} \sum_{i \in S^\bullet} w_i$$

$$\text{s.t. } \sum_{j=1}^{2n} a_{ij}x_j + nw_i \ge d, \quad i \in S^\bullet \tag{15}$$

$$\sum_{j=1}^{2n} a_{ij}x_j \le d-1, \quad i \in S^{\bar\bullet} \tag{16}$$

$$x_j + x_{n+j} \le 1, \quad j \in N \tag{17}$$

$$\sum_{j=1}^{2n} x_j = d \tag{18}$$

$$1 \le d \le n$$

$$\mathbf{x} \in \{0,1\}^{2n}$$

$$\mathbf{w} \in \{0,1\}^{m^\bullet}$$

Now, substituting (18) in (15) and (16), we obtain

$$\sum_{j=1}^{2n} (a_{ij}-1)x_j \ge -nw_i, \quad i \in S^\bullet \tag{19}$$

and

$$\sum_{j=1}^{2n} (1-a_{ij})x_j \ge 1, \quad i \in S^{\bar\bullet}. \tag{20}$$

Let $y_i = -w_i$ and $\alpha_{ij} = a_{ij} - 1$ in (19), and let $\beta_{ij} = 1 - a_{ij}$ in (20). Now, disaggregate the inequality resulting from (19) with respect to each $x_j, j \in \mathcal{N}$, and, via (17), distribute the coefficient of y_i equally among each $x_j, j \in N$. This yields the formulation of ($M_2^\bullet$) above. Following the same transformation steps in reverse order, one obtains (MILP-1$^\bullet$) from ($M_2^\bullet$).

The key difference between the two models, however, is that the integrality on $y_i, i \in S^\bullet$, is relaxed in ($M_2^\bullet$), as shown in the proofs above. Specifically, note that (MILP-1$^\bullet$) is defined in $2n + m^\bullet$ 0–1 integer variables, while ($M_2^\bullet$) involves only $2n$ binary decision variables. In general, n (the number of support features) is a small number, and $m^\bullet \gg n$. Therefore, ($M_2^\bullet$) is much more compact than its counterpart with respect to the hard 0–1 decision variables, hence is expected to generate strong patterns much more efficiently.

Remark 7. When $C^\bullet(p)$ is a singleton with $C^\bullet(p) = \{i\}$, a maximum $C^\bullet(p)$-pattern is nothing but a maximum A_i-pattern. Hence, it is seen that maximum $C^\bullet(p)$-pattern of this paper subsumes the maximum A_i-pattern in the LAD literature [5,12].

Remark 8. In the LAD literature, a pattern p_1 is called a strong $\bullet$ pattern if there is no $\bullet$ pattern p_2 with $C^\bullet(p_1) \subseteq C^\bullet(p_2)$ (e.g., [1,2].) Note, however, that p_1 is coverage-wise not optimal but only Pareto-optimal with respect to $C^\bullet(p_1) \subseteq S^\bullet$. In this paper, we call such a pattern a maximum $C^\bullet(p_1)$-pattern and distinguish it from a strong $\bullet$ pattern that is optimal with respect to the evidential preference.

2.2. Models for Pareto-optimal patterns

Recall that a pattern is prime if deletion of any of its literals makes it a non-pattern. A strong pattern that is also Pareto-optimal with respect to this simplicity criterion is called a strong prime pattern. To develop a model for strong prime $\bullet$ patterns, let us select a real number $\omega \in (0, \frac{1}{n+1}]$ (where n is the number of support features), and consider the following model named ($M_{sp}^\bullet$).

$$c_{sp} = \max_{\mathbf{x},\mathbf{y}} \sum_{i \in S^\bullet} y_i - \omega \sum_{j \in \mathcal{N}} x_j$$

$$\text{s.t. (7)–(10)}$$

Theorem 9. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be an optimal solution of ($M_{sp}^\bullet$) for $\omega \in (0, \frac{1}{n+1}]$. Then, p defined as (6) forms a strong prime $\bullet$ pattern with coverage c_{sp}^*.*

Before proving this result, let us note that a strong prime $\bullet$ pattern can be formed from the solution $(\mathbf{x}^\dagger, \mathbf{y}^\dagger)$ obtained by first solving ($M_s^\bullet$) to get c_s^* and next solving the following MILP model:

$$c = \min_{\mathbf{x},\mathbf{y}} \sum_{j \in \mathcal{N}} x_j$$

$$\text{s.t. (7)–(10)}$$

$$\sum_{i \in S^\bullet} y_i \ge c_s^*$$

Proof. The primeness of the pattern formed by $(\mathbf{x}^*, \mathbf{y}^*)$ of $(\mathrm{M}^{\bullet}_{sp})$ is obvious, hence we will only show that the pattern is also a strong pattern. Toward this end, suppose that $(\mathbf{x}^*, \mathbf{y}^*)$ does not form a strong pattern; that is,

$$\sum_{i\in S^{\bullet}} y_i^{\dagger} > \sum_{i\in S^{\bullet}} y_i^*,$$

and the integrality of $\mathbf{y}^{\dagger}$ and $\mathbf{y}^*$ yields

$$\sum_{i\in S^{\bullet}} y_i^{\dagger} \geq \sum_{i\in S^{\bullet}} y_i^* + 1. \tag{21}$$

Note that $x_j^{\dagger} + x_{n+j}^{\dagger} \leq 1$ and $x_j^* + x_{n+j}^* \leq 1$ for all $j \in N$, hence we have $0 < \sum_{j\in\mathcal{N}} x_j^{\dagger} \leq n$ and $0 < \sum_{j\in\mathcal{N}} x_j^* \leq n$. Now, let ω be a real number from the interval $(0, \frac{1}{n+1}]$ and multiply the three sides in these three-part inequalities by ω. This yields

$$0 < \omega \sum_{j\in\mathcal{N}} x_j^{\dagger} \leq \omega n \leq \frac{n}{n+1} < 1 \quad \text{and} \quad 0 < \omega \sum_{j\in\mathcal{N}} x_j^* \leq \omega n \leq \frac{n}{n+1} < 1,$$

and we have

$$\omega \sum_{j\in\mathcal{N}} x_j^{\dagger} - \omega \sum_{j\in\mathcal{N}} x_j^* < 1. \tag{22}$$

Putting (21) and (22) together, we have

$$\sum_{i\in S^{\bullet}} y_i^{\dagger} \geq \sum_{i\in S^{\bullet}} y_i^* + 1 > \sum_{i\in S^{\bullet}} y_i^* + \omega \sum_{j\in\mathcal{N}} x_j^{\dagger} - \omega \sum_{j\in\mathcal{N}} x_j^* \Longleftrightarrow \sum_{i\in S^{\bullet}} y_i^{\dagger} - \omega \sum_{j\in\mathcal{N}} x_j^{\dagger} > \sum_{i\in S^{\bullet}} y_i^* - \omega \sum_{j\in\mathcal{N}} x_j^*.$$

Now, note that $(\mathbf{x}^{\dagger}, \mathbf{y}^{\dagger})$ satisfies all constraints in (7)–(10) of $(\mathrm{M}^{\bullet}_{sp})$. Thus, the last inequality contradicts the fact that $(\mathbf{x}^*, \mathbf{y}^*)$ is an optimal solution of $(\mathrm{M}^{\bullet}_{sp})$, and this completes the proof. □

Recall that a spanned pattern is selectivity-wise an optimal pattern. We can see from this that spanned patterns are complex and specific patterns. A strong pattern that is also Pareto-optimal with respect to this selectivity measure is called a strong spanned pattern. Based upon the basic idea presented for $(\mathrm{M}^{\bullet}_{sp})$, we can develop an MILP model for strong spanned • patterns. For the purpose, let us select a real number $\omega \in [-\frac{1}{n+1}, 0)$ and consider the following model named $(\mathrm{M}^{\bullet}_{ss})$.

$$c_{ss} = \max_{\mathbf{x},\mathbf{y}} \sum_{i\in S^{\bullet}} y_i - \omega \sum_{j\in\mathcal{N}} x_j$$
$$\text{s.t. (7)–(10)}$$

Corollary 10. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be an optimal solution of $(\mathrm{M}^{\bullet}_{ss})$ for $\omega \in [-\frac{1}{n+1}, 0)$. Then, p defined as (6) forms a strong spanned • pattern with coverage c^*_{ss}.*

Proof. Follow the steps of the proof for Theorem 9. □

Last, let us modify $(\mathrm{M}^{\bullet}_{m})$ of the last subsection for maximum $C^{\bullet}(p)$-patterns and obtain the following MILP model named $(\mathrm{M}^{\bullet}_{m,s/p})$.

$$c_{m,s/p} = \max_{\mathbf{x},\mathbf{y}} \sum_{i\in S^{\bullet}\setminus C^{\bullet}(p)} y_i - \omega \sum_{j\in J_p} x_j$$
$$\text{s.t. (11)–(14)}$$

The proof for the following theorem is similar to the one for Theorem 9.

Theorem 11. *Let $(\mathbf{x}^*, \mathbf{y}^*)$ be an optimal solution of $(\mathrm{M}^{\bullet}_{m,s/p})$ for $\omega \in (0, \frac{1}{|J_p|+1}]$ ($\omega \in [-\frac{1}{|J_p|+1}, 0)$). Then, p defined as (6) forms a maximum prime $C^{\bullet}(p)$-pattern (a maximum spanned $C^{\bullet}(p)$-pattern) with coverage $c^*_{m,s/p} + |C^{\bullet}(p)|$.*

3. Numerical studies

This section demonstrates the efficiency of the new compact MILP models and investigate different utilities of strong prime patterns and strong spanned patterns in enhancing the overall classification accuracy of a LAD decision theory. For these experiments, we used six well-solved machine learning datasets from [11] in Table 1. The computing platform used was a Linux PC with an Intel i7 3.4 GHz 8-core processor chip with 12 Gb of memory. We used Gurobi Optimizer 4.5.2 [8] for solving the MILP instances generated during these experiments to optimality.

Table 1
Datasets used.

Dataset (abbreviation)	Number of observations (+, −)
Boston housing (`housing`)	506 (260 with income ≥$21 K, 246 else)
BUPA liver disorder (`liver`)	345 (145 selector 1, 200 selector 2)
Cleveland heart disease (`heart`)	297 (137 disease, 160 no disease)
Credit card scoring (`credit`)	653 (296 approvals, 357 denials)
Pima Indian diabetes (`diabetes`)	768 (268 diabetes, 500 no disease)
Wisconsin breast cancer (`wbc`)	683 (239 malignant, 444 benign)

3.1. Utility of compact MILP pattern generation models

The MILP models of the previous section involve a much smaller number of 0–1 decision variables than their counterparts from [12], hence can generate useful LAD patterns more efficiently. To demonstrate this utility of the new pattern generation models, we used ($M_s^\bullet$) above and (MILP-1$^\bullet$) from [12] to generate strong patterns for the six datasets in Table 1 and compared their performance. Recall that ($M_s^\bullet$) involves $2n$ 0–1 variables, while its counterpart has $2n + m^\bullet$ 0–1 variables, where n is the number of support features and $m^\bullet$ is the number of $\bullet$ observations.

Patterns are discovered from training data, hence the efficiency of a pattern generation model is better illustrated if a larger number of data is used for training. For these experiments, therefore, we adopted 10-fold cross-validation experiments with a random split of the dataset under analysis into 10 equal size and mutually disjoint partitions. After a training dataset was formed by combining 9 of the 10 partitions, we followed the standard data binarization and support feature selection steps for LAD from [7] and used each of the two MILP models in turn in **procedure** 1 `pattern generation` below to generate strong patterns. This process was repeated a total of 10 times for 10 different ways of forming a training dataset.

Table 2 reports the average results and the corresponding standard deviations in 'average ± standard deviation' format. Specifically, the table provides the number of the 0–1 decision variables in the first (hence the largest) (MILP-1$^\bullet$) and ($M_s^\bullet$) instances generated and then gives time in CPU seconds in which the two models generated a complete set of strong + and strong − patterns for each of the six datasets. On the `diabetes` data, we ran (MILP-1$^\bullet$) only once, as we deemed the model required too much time of about 24.7 CPU h (=1700+87,132 CPU s) for generating one set of patterns. For a direct comparison, the last two legends of Table 2 report the results by ($M_s^\bullet$) on the same + and − training `diabetes` data for which (MILP-1$^\bullet$) generated the strong + and − patterns in Table 2, respectively.

In summary, Table 2 shows that ($M_s^\bullet$) generated strong patterns up to more than 2 orders of magnitude more efficiently in these experiments. Briefly, this efficiency of ($M_s^\bullet$) over (MILP-1$^\bullet$) owes to an order of magnitude fewer 0–1 variables in the ($M_s^\bullet$) instances generated during these experiments, and this illustrates the usefulness of the compact MILP models of this paper well.

3.2. Utility of strong prime patterns and strong spanned patterns

In supervised learning, a simpler rule is believed to possess a better generalization capability and classify new observations more accurately (e.g., [4,10].) As the simplest of all strong patterns, therefore, strong prime patterns are expected to help improve the classification accuracy (or the generalization capability) of a LAD decision theory. On the other hand, a strong spanned pattern is the most complex (involving the largest number of literals) of all strong patterns, hence is not likely to react to noisy observations. Therefore, strong spanned patterns can help a LAD decision theory in reducing the number of misclassified decisions. In this subsection, we investigate how these two 'perhaps the most' useful Pareto-optimal patterns contribute differently to enhancing the overall classification capability of a LAD decision theory.

For these experiments, we used (M_{sp}) and (M_{ss}) and tested their performance in 10 repeated runs of 50–50 holdout experiments with a random split of a dataset into two equal halves, one for training and the other for testing. For a realistic treatment of a real-life test setting, we used only the training data for deriving cutpoints for data binarization and for support

procedure 1 `pattern generation`

input: training data, support features, MILP model for pattern generation
output: a set of + and − patterns (P^+ and P^-, respectively)
for $\bullet \in \{+, -\}$ **do**
  set $P^\bullet = \emptyset$
  while $S^\bullet \neq \emptyset$ **do**
    formulate and solve an instance of MILP.
    form a pattern p from the solution obtained.
    $P^\bullet \leftarrow P^\bullet \cup \{p\}$
    $S^\bullet \leftarrow S^\bullet \setminus \{i \in S^\bullet : A_i$ is covered by $p\}$
  end while
end for

Table 2
Pattern generation by (MILP-1) and (M_s).

Dataset	Class	(MILP-1)		(M_s)	
		0–1 variables[a]	CPU seconds[b]	0–1 variables[a]	CPU seconds[b]
`Housing`	+	265.5 ± 3.5	29.1 ± 4.7	34.2 ± 2.0	11.2 ± 3.4
	−	258.3 ± 4.1	45.7 ± 9.3	34.2 ± 2.0	18.2 ± 5.3
`Liver`	+	219.6 ± 1.6	230.8 ± 57.4	39.6 ± 1.6	49.1 ± 14.3
	−	170.1 ± 2.6	84.4 ± 11.7	39.6 ± 1.6	27.3 ± 9.4
`Heart`	+	146.5 ± 2.2	16.6 ± 2.8	23.2 ± 1.0	3.1 ± 0.7
	−	167.2 ± 1.0	21.1 ± 3.1	23.2 ± 1.0	3.7 ± 1.1
`Credit`	+	354.1 ± 3.6	5,676.7 ± 4,713.6	32.8 ± 1.9	47.9 ± 19.5
	−	299.2 ± 3.3	1,028.1 ± 465.1	32.8 ± 1.9	44.1 ± 11.6
`Diabetes`	+	286[c]	1,700[c]	42.8 ± 1.4	506.2 ± 109.0
	−	494[d]	87,132[d]	42.8 ± 1.4	557.6 ± 118.9
`wbc`	+	238.5 ± 3.0	6.1 ± 1.4	23.4 ± 1.3	2.9 ± 1.3
	−	423.0 ± 1.7	2.0 ± 1.1	23.4 ± 1.3	0.6 ± 0.7

Results are provided in 'average ± standard deviation' format.
[a] Number of 0–1 variables in the first instance of MILP generated.
[b] Pattern generation time by the MILP model.
[c] (M_s) required 518 CPU s for the same + `diabetes` data (Due to excessive time required by (MILP-1), here we report the time of a single run).
[d] (M_s) required 400 CPU s for the same − `diabetes` data (Due to excessive time required by (MILP-1), here we report the time of a single run).

true class / decision	+	−
+	sensitivity	type I error
−	type II error	specificity
neither	type 0 error: unclassified error	

Fig. 1. Types of accuracy and errors in decisions.

feature selection; recall that the testing data are future observations that are not available during the training stage. Next, we applied **procedure** 1 `pattern generation` with (M_{sp}) and (M_{ss}) to generate a set of strong prime patterns and a set of strong spanned patterns, respectively, and formed two LAD decision theories, one comprised solely of the strong prime patterns and the other of the strong spanned patterns. As in [7], we formed a LAD decision theory as the difference between the weighted averages of the + and − patterns using the prevalence $\frac{c^{\bullet}(p)}{m^{\bullet}}$ of a • pattern p for its weight. For an objective assessment of different utilities of the two types of pattern, we used the perfect training philosophy and did not employ any heuristic measure, such as the prevalence of a pattern, for pre-sorting or selecting which patterns to use in forming LAD theories. Next, we directly applied the cutpoints generated during the training stage to binarize the testing data and applied the LAD decision theory formed to classify the binarized testing data. Finally, we counted the number of correct classifications and occurrences of each of the three types of decision error in Fig. 1.

Table 3 summarizes information on the number of patterns and the degree of patterns generated by the two MILP models, and Table 4 summarizes information on the coverage of these patterns on the training data. Again, all results in these tables are provided in format 'average ± standard deviation' format of the 10 results from the 10 repeated runs of 50–50 holdout experiments, and 'Min', 'Avg' and 'Max' in these tables refer to the minimum, average and maximum of the 10 results, respectively.

In Table 3, we first note that (M_{sp}) and (M_{ss}) generated about the same number of patterns for the six datasets. In terms of the degree, however, we note a striking difference between (M_{sp}) and (M_{ss}) patterns, which originates from the difference in nature between the two types of Pareto-optimal pattern. In summary, we see in Table 3 that the strong spanned patterns are about 3 to 4 times more complex than the strong prime patterns. Next, Table 4 shows that the minimum and the average coverages of the strong prime patterns are much superior to those of the strong spanned patterns.

Table 5 compares the testing performance of the two types of Pareto-optimal pattern and shows that strong prime patterns are superior in terms of the number of accurate decisions and unclassified errors while strong spanned patterns are better in terms of the number of misclassification decisions. To see this clearly, compare the testing accuracies of the Pareto-optimal patterns on 'harder-to-classify' `liver`, `heart` and `diabetes` data and note that the difference in testing accuracy is between 8% (=75.1%–67.1% on − `diabetes` data) and 15.1% (=64.4%–49.3% on + `liver` data) in favor of the strong prime

Table 3
Number and degree of patterns generated by (M_{sp}) and (M_{ss}).

Dataset	Class	Strong prime patterns by (M_{sp})				Strong spanned patterns by (M_{ss})			
		Number	Min	Avg	Max	Number	Min	Avg	Max
Housing	+	13.9 ± 2.0	2.1 ± 0.6	3.8 ± 0.2	5.8 ± 0.4	14.0 ± 2.1	6.4 ± 1.6	13.3 ± 2.1	17.8 ± 1.8
	−	14.3 ± 2.3	1.9 ± 0.6	3.7 ± 0.3	5.3 ± 0.5	14.1 ± 2.1	6.0 ± 1.7	13.1 ± 1.7	17.8 ± 1.8
Liver	+	22.1 ± 2.2	2.0 ± 0.5	4.0 ± 0.2	5.8 ± 0.4	22.5 ± 2.2	7.2 ± 0.8	15.0 ± 1.0	19.8 ± 1.9
	−	21.0 ± 2.4	2.7 ± 0.5	4.3 ± 0.3	6.7 ± 0.8	21.5 ± 2.3	9.2 ± 1.9	16.2 ± 1.3	19.8 ± 1.9
Heart	+	13.9 ± 1.9	2.6 ± 0.5	3.8 ± 0.2	5.1 ± 0.3	13.8 ± 1.8	3.7 ± 0.8	9.0 ± 1.2	12.2 ± 1.4
	−	13.3 ± 1.4	2.9 ± 0.3	4.0 ± 0.2	5.2 ± 0.4	13.6 ± 1.4	4.3 ± 0.7	8.6 ± 1.1	12.2 ± 1.4
Credit	+	21.4 ± 1.6	2.3 ± 0.5	4.5 ± 0.2	6.6 ± 1.0	21.3 ± 1.7	3.6 ± 1.2	11.5 ± 1.6	17.5 ± 1.7
	−	22.5 ± 1.6	3.2 ± 0.4	4.7 ± 0.2	6.4 ± 0.7	22.2 ± 1.5	4.7 ± 0.7	12.2 ± 1.2	17.5 ± 1.7
Diabetes	+	33.0 ± 4.1	2.9 ± 0.3	5.0 ± 0.2	7.6 ± 0.7	32.1 ± 4.7	7.6 ± 1.8	15.7 ± 1.9	21.3 ± 3.0
	−	33.6 ± 2.8	2.9 ± 0.6	4.8 ± 0.2	7.4 ± 0.8	32.7 ± 3.1	7.6 ± 1.6	14.8 ± 2.0	21.3 ± 3.0
wbc	+	8.6 ± 1.4	1.8 ± 0.4	2.8 ± 0.2	4.2 ± 0.6	9.0 ± 1.6	2.8 ± 0.6	6.5 ± 1.1	10.0 ± 1.6
	−	6.9 ± 1.0	2.1 ± 0.3	3.7 ± 0.3	5.2 ± 0.6	6.9 ± 1.3	4.6 ± 0.8	8.1 ± 1.0	10.1 ± 1.2

Results are provided in 'average ± standard deviation' format.

Table 4
Coverage of patterns generated by (M_{sp}) and (M_{ss}) on 50% training data.

Dataset	Class	Strong prime patterns by (M_{sp})			Strong spanned patterns by (M_{ss})		
		Min	Avg	Max	Min	Avg	Max
Housing	+	1.9 ± 1.0	20.7 ± 3.5	74.1 ± 8.7	1.0 ± 0.0	12.0 ± 2.3	74.1 ± 8.7
	−	1.5 ± 0.8	16.4 ± 2.3	56.7 ± 12.8	1.0 ± 0.0	10.8 ± 2.2	56.7 ± 12.8
Liver	+	1.4 ± 0.5	6.8 ± 0.9	16.6 ± 2.7	1.0 ± 0.0	4.9 ± 0.7	16.6 ± 2.7
	−	1.1 ± 0.3	4.7 ± 0.7	14.8 ± 3.2	1.0 ± 0.0	3.6 ± 0.5	14.8 ± 3.2
Heart	+	1.5 ± 0.7	10.3 ± 1.6	31.3 ± 6.1	1.0 ± 0.0	5.8 ± 0.9	31.3 ± 6.1
	−	1.5 ± 0.7	11.4 ± 2.5	33.6 ± 4.0	1.0 ± 0.0	7.1 ± 1.3	33.6 ± 4.0
Credit	+	1.5 ± 0.7	17.1 ± 1.9	57.9 ± 12.1	1.0 ± 0.0	11.5 ± 1.7	57.9 ± 12.1
	−	1.3 ± 0.5	12.9 ± 1.1	47.7 ± 5.8	1.0 ± 0.0	8.8 ± 0.7	47.7 ± 5.8
Diabetes	+	1.0 ± 0.0	6.1 ± 0.8	19.5 ± 3.5	1.0 ± 0.0	4.8 ± 0.8	19.5 ± 3.5
	−	1.2 ± 0.4	13.7 ± 2.4	67.1 ± 8.1	1.0 ± 0.0	10.3 ± 1.7	67.1 ± 8.1
wbc	+	4.6 ± 1.9	32.0 ± 6.9	72.0 ± 9.4	1.4 ± 1.3	16.1 ± 3.1	72.0 ± 9.4
	−	1.0 ± 0.0	84.7 ± 26.1	205.0 ± 4.5	1.1 ± 0.3	50.8 ± 20.2	205.0 ± 4.5

All results are provided in 'average ± standard deviation' format.

patterns. When their testing accuracies are normalized with respect to the lower of the two rates on each type of data of the three aforementioned datasets, the difference is magnified and comes out to be between 11.9% ($=\frac{75.1\%-67.1\%}{67.1\%} \times 100\%$ on − `diabetes` data) and 30.6% ($=\frac{64.4\%-49.3\%}{49.3\%} \times 100\%$ on + `liver` data), in favor of the strong prime patterns.

When the two Pareto-optimal patterns are compared in terms of the number of misclassified decisions, however, the results in Table 5 favor the strong spanned patterns. On the three aforementioned harder-to-classify datasets, for example, the difference in the number of misclassification errors comes out to be between 3.9% (on − `diabetes` data) and 8.9% (on − `liver` data) in a direct comparison of the numbers and between 25.7% ($=\frac{19.1\%-15.2\%}{15.2\%} \times 100\%$ on − `diabetes` data) and 39.3% ($=\frac{18.8\%-13.5\%}{13.5\%} \times 100\%$ on − `heart` data) when the normalized results are compared, in favor of the strong spanned patterns. We note that this benefit of the strong spanned patterns is accompanied by a high cost of unclassified decisions, though.

In summary, we believe that the experiments in this subsection confirm that strong prime and strong spanned patterns have different and specific utilities in supervised learning; specifically, strong prime patterns generalize better on new observations and increase the sensitivity and the specificity of decisions while strong spanned patterns reduce the risk of making misclassification errors. This is intuitive from the definitions of these useful patterns, and is now supported also by the experimental evidence provided in this subsection.

4. Concluding remarks

In view of the fact that term-enumerative methods can be quite limited in generating useful patterns, Ryoo and Jang [12] introduced the notion of MILP-based pattern generation and presented the first generation of MILP models for generating patterns that are optimal or Pareto-optimal with respect to the coverage (evidential), simplicity and selectivity preferences.

Using the evidential preference, we developed in this paper a new MILP model for strong patterns that involves a much smaller number of hard 0–1 integer variables than the one presented in [12]. Next, we used the new MILP model to develop compact MILP models for generating strong prime patterns and strong spanned patterns and then developed

Table 5
Classification results of patterns generated by (M_{sp}) and (M_{ss}) on 50% testing data.

Dataset	Class	Number of data	Strong prime patterns by (M_{sp})			Strong spanned patterns by (M_{ss})		
			Accurate decisions	Errors		Accurate decisions	Errors	
				Type I/II	Type 0		Type I/II	Type 0
Housing	+	129 (100%)	105.1 ± 4.9 (81.5%)	21.0 ± 4.1 (16.3%)	2.9 ± 3.1 (2.2%)	99.2 ± 6.2 (76.9%)	17.4 ± 5.5 (13.5%)	12.4 ± 5.0 (9.6%)
	−	125 (100%)	102.0 ± 5.8 (81.6%)	21.1 ± 5.8 (16.9%)	1.9 ± 1.6 (1.5%)	94.1 ± 7.8 (75.3%)	15.5 ± 5.4 (12.4%)	15.4 ± 4.7 (12.3%)
Liver	+	100 (100%)	64.4 ± 4.5 (64.4%)	28.8 ± 3.4 (28.8%)	6.8 ± 2.8 (6.8%)	49.3 ± 6.3 (49.3%)	21.7 ± 3.5 (21.7%)	29.0 ± 5.5 (29.0%)
	−	73 (100%)	40.2 ± 4.9 (55.1%)	27.4 ± 5.7 (37.5%)	5.4 ± 2.4 (7.4%)	31.2 ± 5.4 (42.7%)	20.9 ± 4.8 (28.6%)	20.9 ± 5.1 (28.6%)
Heart	+	69 (100%)	48.8 ± 3.2 (70.7%)	18.1 ± 2.7 (26.2%)	2.1 ± 1.9 (3.0%)	40.8 ± 4.3 (59.1%)	13.2 ± 3.0 (19.1%)	15.0 ± 4.3 (21.7%)
	−	80 (100%)	62.6 ± 5.3 (78.3%)	15.0 ± 5.8 (18.8%)	2.4 ± 1.8 (3.0%)	53.7 ± 7.1 (67.1%)	10.8 ± 4.5 (13.5%)	15.5 ± 4.9 (19.4%)
Credit	+	179 (100%)	144.9 ± 7.4 (80.9%)	27.9 ± 6.1 (15.6%)	6.2 ± 2.5 (3.5%)	134.3 ± 6.4 (75.0%)	20.2 ± 3.7 (11.3%)	24.5 ± 3.7 (13.7%)
	−	148 (100%)	115.2 ± 4.7 (77.8%)	25.4 ± 4.5 (17.2%)	7.4 ± 3.9 (5.0%)	105.2 ± 7.6 (71.1%)	20.0 ± 3.6 (13.5%)	22.8 ± 7.3 (15.4%)
Diabetes	+	134 (100%)	68.7 ± 7.7 (51.3%)	55.8 ± 5.8 (41.6%)	9.5 ± 3.6 (7.1%)	54.8 ± 5.5 (40.9%)	44.0 ± 5.3 (32.8%)	35.2 ± 6.0 (26.3%)
	−	250 (100%)	187.7 ± 8.6 (75.1%)	47.8 ± 5.3 (19.1%)	14.5 ± 5.9 (5.8%)	167.7 ± 7.3 (67.1%)	38.0 ± 7.5 (15.2%)	44.3 ± 7.1 (17.7%)
wbc	+	120 (100%)	106.9 ± 4.2 (89.1%)	10.9 ± 5.1 (9.1%)	2.2 ± 1.9 (1.8%)	103.2 ± 3.3 (86.0%)	7.1 ± 4.2 (5.9%)	9.7 ± 2.0 (8.1%)
	−	222 (100%)	213.8 ± 2.9 (96.3%)	7.8 ± 3.1 (3.5%)	0.4 ± 0.7 (0.2%)	211.3 ± 3.4 (95.2%)	7.4 ± 3.9 (3.3%)	3.3 ± 2.7 (1.5%)

Results are provided in 'average ± standard deviation' format.
Average testing accuracy and error rate (in per cent) are provided in parentheses for reference.

new MILP models for maximum $C^\bullet(p)$-patterns maximum prime $C^\bullet(p)$-patterns, and maximum spanned $C^\bullet(p)$-patterns. With numerical experiments on six benchmark machine learning datasets, we showed the efficiency of the new compact pattern generation models over their counterparts from [12]. We also demonstrated different and specific utilities of the strong prime and strong spanned patterns in a LAD decision theory; specifically, strong prime patterns help increase the sensitivity and the specificity of decisions, while strong spanned patterns help reduce the risk of misclassification errors. This is intuitive from the definitions of these useful patterns and is now supported also by the experimental evidence provided in this paper.

References

[1] G. Alexe, S. Alexe, T. Bonates, A. Kogan, Logical analysis of data — the vision of Peter L. Hammer, Ann. Math. Artif. Intell. 49 (2007) 265–312.
[2] G. Alexe, S. Alexe, P. Hammer, A. Kogan, Comprehensive vs. comprehensible classifiers in logical analysis of data, Discrete Appl. Math. 156 (6) (2008) 870–882.
[3] G. Alexe, P. Hammer, Spanned patterns for the logical analysis of data, Discrete Math. 154 (7) (2006) 1039–1049.
[4] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth, Occam's razor, Inform. Process. Lett. 24 (1987) 377–380.
[5] T. Bonates, P. Hammer, A. Kogan, Maximum patterns in datasets, Discrete Appl. Math. 156 (6) (2008) 846–861.
[6] E. Boros, P. Hammer, T. Ibaraki, A. Kogan, Logical analysis of numerical data, Math. Program. 79 (1997) 163–190.
[7] E. Boros, P. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik, An implementation of logical analysis of data, IEEE Trans. Knowl. Data Eng. 12 (2000) 292–306.
[8] Gurobi Optimization, Gurobi Optimizer Reference Manual Version 4.0, Gurobi Optimization, Houston, Texas, November 2010.
[9] P. Hammer, A. Kogan, B. Simeone, S. Szedmak, Pareto-optimal patterns in logical analysis of data, Discrete Appl. Math. 144 (2004) 79–102.
[10] R. Holte, Very simple classification rules perform well on most commonly used datasets, Mach. Learn. 11 (1993) 63–91.
[11] P. Murphy, D. Aha, Uci repository of machine learning databases: readable data repository, Department of Computer Science, University of California at Irvine, CA, 1994. Available from World Wide Web: http://www.ics.uci.edu/~mlearn/MLRepository.html.
[12] H. Ryoo, I.-Y. Jang, Milp approach to pattern generation in logical analysis of data, Discrete Appl. Math. 157 (4) (2009) 749–761.